Equidistribution of Horocyclic Flows on Complete Hyperbolic Surfaces of Finite Area

John H. Hubbard and Robyn L. Miller
Cornell University

March 2, 2008

Abstract

We provide a self-contained, accessible introduction to Ratner's Equidistribution Theorem in the special case of horocyclic flow on a complete hyperbolic surface of finite area. This equidistribution result was first obtained in the early 1980s by Dani and Smillie [DS84] and later reappeared as an illustrative special case [Rat92] of Ratner's work [Rat91-Rat94] on the equidistribution of unipotent flows in homogeneous spaces. We also prove an interesting probabilistic result due to Breuillard: on the modular surface an arbitrary uncentered random walk on the horocycle through almost any point will fail to equidistribute, even though the horocycles are themselves equidistributed [Bre05]. In many aspects of this exposition we are indebted to Bekka and Mayer's more ambitious survey [BM00], *Ergodic Theory and Topological Dynamics for Group Actions on Homogeneous Spaces.*

1 Horocycle flow on hyperbolic surfaces

Let X be a complete hyperbolic surface, perhaps the hyperbolic plane H, and let $\mathbf{X}$ denote the unit tangent bundle $T^1(X)$ to X (and $\mathbf{H} = T^1H$). There are three flows on $\mathbf{X}$ which will concern us here. They are realized by three cars, as represented in Figure 1.

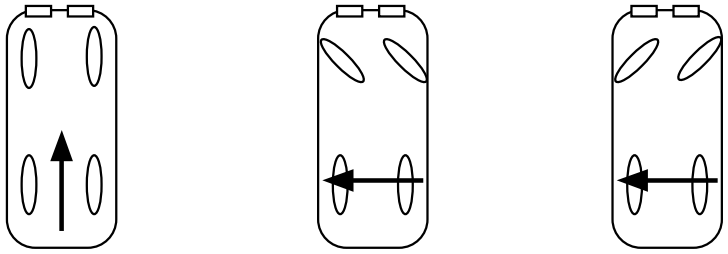

Figure 1: Driving the cars above leads to geodesic flow, positive horocycle flow and negative horocycle flow respectively

The cars all have their steering wheels locked in position: the first car drives straight ahead, the second one steers to the left so as to follow a path of geodesic curvature 1, and the third steers to the right, also following a path of geodesic curvature 1. All three cars have an arrow painted on the roof, centered at the rear axle; for the first the arrow points straight ahead, and for the other two it points sideways – in the direction towards which the car is steering for the second car and in the opposite direction for the third.

The flows at time $t \in \mathbb{R}$ starting at a point $\mathbf{x} = (x, \xi) \in \mathbf{X}$ are defined as follows:

1. *The geodesic flow*: put the first car on X with the arrow pointing in the direction of ξ, and drive a distance t. The point of arrival, with the arrow on the car at that point, will be denoted $\mathbf{x}g(t)$;

2. *The positive horocyclic flow*: put the second car on X with the arrow pointing in the direction of ξ, and drive a distance t. The point of arrival, with the arrow on the car at that point, will be denoted $\mathbf{x}u_+(t)$;

3. *The negative horocyclic flow*: put the third car on X with the arrow pointing in the direction of ξ, and drive a distance t. The point of arrival, with the arrow on the car at that point, will be denoted $\mathbf{x}u_-(t)$.

We will see when we translate to matrices why it is convenient to write the flows as *right* actions.

The trajectories followed by these cars are represented in Figure 2.

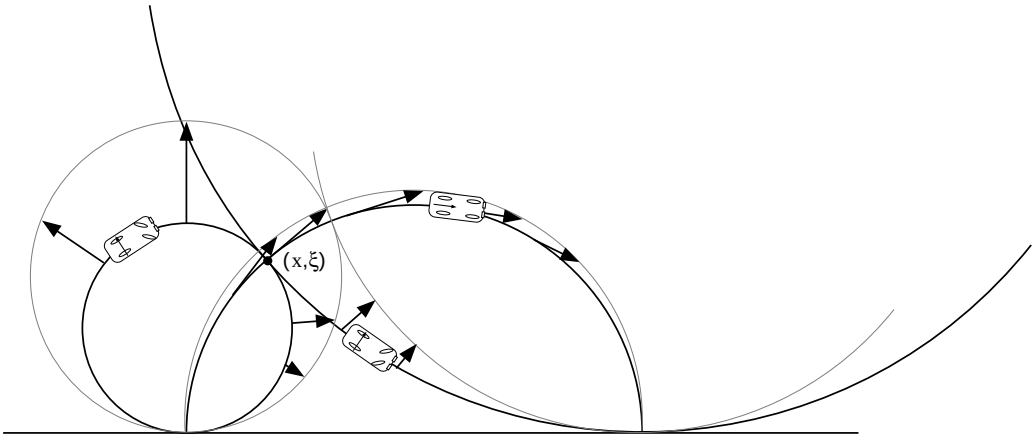

Figure 2: In the upper half-plane model of the hyperbolic plane, the geodesic passing through (x, ξ) is the semicircle perpendicular to the real axis and tangent at x to ξ. One should remember that it is not a curve in $\mathbb{H}$, but rather a curve in $T^1(\mathbb{H})$ and carries its velocity vector with it. From the point (x, ξ), the positive horocycle is the circle tangent to the real axis at the endpoint of the geodesic above and perpendicular to ξ at x, whereas the negative horocycle flow is the circle tangent to the real axis at the origin of the geodesic, and still perpendicular to ξ at x. We have drawn our tinkertoys driving along them.

2 Translation to Matrices

In less picturesque language (more formal, not more accurate), you can identify $\mathbf{X}$ with $\Gamma \setminus \mathrm{PSL}_2 \mathbb{R}$ for some Fuchsian group Γ.

1. The geodesic flow of the point represented by $g \in \mathrm{PSL}_2 \mathbb{R}$ is
$$t \mapsto g \begin{pmatrix} e^{\frac{t}{2}} & 0 \\ 0 & e^{-\frac{t}{2}} \end{pmatrix};$$
2. The positive horocyclic flow of the point represented by $g \in \mathrm{PSL}_2 \mathbb{R}$ is
$$t \mapsto g \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix};$$
3. The negative horocyclic flow of the point represented by $g \in \mathrm{PSL}_2 \mathbb{R}$ is
$$t \mapsto g \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix};$$

The standard left action of $\mathrm{PSL}_2 \mathbb{R}$ on H, which lifts by the derivative to a left action on $\mathbf{H}$ is given by

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot z = \frac{az+b}{cz+d} \quad \text{lifting to} \quad \begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot (z, \xi) = \left(\frac{az+b}{cz+d}, \frac{\xi}{(cz+d)^2} \right) \quad (1)$$

We can then identify $\mathrm{PSL}_2 \mathbb{R}$ to $\mathbf{H}$ by choosing $\mathbf{x_0} = (i, i) \in \mathbf{H}$ and setting $\Phi : \mathrm{PSL}_2 \mathbb{R} \to \mathbf{H}$ to be

$$\Phi \begin{pmatrix} a & b \\ c & d \end{pmatrix} := \begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \mathbf{x_0} = \left(\frac{ai+b}{ci+d}, \frac{i}{(ci+d)^2} \right)$$

Since

$$\Phi(\gamma A) = (\gamma A) \cdot \mathbf{x_0} = \gamma \cdot (A \cdot \mathbf{x_0}) = \gamma \cdot \Phi(A)$$

we see that Φ induces a diffeomorphism $\Phi_\Gamma : \Gamma \setminus \mathrm{PSL}_2 \mathbb{R} \to \Gamma \setminus \mathbf{X}$.

The left action above does *not* induce an action of $\mathrm{PSL}_2 \mathbb{R}$ on $\Gamma \setminus \mathbf{X}$, but there is an action on the right given by

$$\Phi(A) * B = \Phi(AB)$$

For $\gamma \in \Gamma$ we have

$$\Phi_\Gamma(A) * B = \Phi_\Gamma(AB) = \Phi_\Gamma(\gamma AB) = \gamma \cdot \Phi_\Gamma(AB) = \gamma \cdot (\Phi_\Gamma(A) * B)$$

so the action is well defined on $\mathbf{X}$. All three flows are special cases, eg. write

$$G^t = \begin{pmatrix} e^{\frac{t}{2}} & 0 \\ 0 & e^{-\frac{t}{2}} \end{pmatrix}, \quad U_+^t = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, \quad U_-^t = \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}$$

and name the corresponding one-parameter subgroups

$$G = \{G^t |\ t \in \mathbb{R}\}, \quad U_+ = \{U_+^t |\ t \in \mathbb{R}\}, \quad U_- = \{U_-^t |\ t \in \mathbb{R}\}.$$

Then

$$\mathbf{x}g(t) = \mathbf{x} * G^t, \quad \mathbf{x}u_+(t) = \mathbf{x} * U_+^t, \quad \mathbf{x}u_-(t) = \mathbf{x} * U_-^t$$

Let us check these. By naturality we see that for all $A \in \mathrm{PSL}_2\,\mathbb{R}$ we have

$$(A \cdot \mathbf{x_0})g(t) = A \cdot (\mathbf{x_0}g(t)),\ (A \cdot \mathbf{x_0})u_+(t) = A \cdot (\mathbf{x_0}u_+(t)),\ (A \cdot \mathbf{x_0})u_-(t) = A \cdot (\mathbf{x_0}u_-(t))$$

and, moreover

$$\Phi(G^t) = \mathbf{x}_0 g(t), \quad \Phi(U_+^t) = \mathbf{x}_0 u_+(t), \quad \Phi(U_-^t) = \mathbf{x}_0 u_-(t)$$

so

$$\Phi(AG^t) = (AG^t) \cdot \mathbf{x_0} = A \cdot (G^t \cdot \mathbf{x_0}) = A \cdot (\mathbf{x_0}g(t)) = (A \cdot \mathbf{x_0})g(t) = \Phi(A)g(t)$$

and the argument for u_+ and u_- is identical.

Left multiplication by G^t, U_+^t and $U_-(t)$ also give flows on $\mathbf{H}$; probably easier to understand than the geodesic and horocycle flows. For instance, left action by U_+^t corresponds to translating a point and vector to the right by t. But these actions do not commute with the action of Γ and hence induce nothing on $\mathbf{X}$.

Since $\mathrm{PSL}_2\,\mathbb{R}$ is unimodular, it has a Haar measure, invariant under both left and right translation, and unique up to multiples. Since $\mathrm{PSL}_2\,\mathbb{R}$ is not compact, there is no natural normalization. Denote by ω the corresponding measure on $\mathbf{H}$; if $\mathbf{X} = \Gamma\backslash\mathbf{H}$ is of finite volume, we will denote by $\omega_{\mathbf{X}}$ the corresponding measure normalized so that $\omega_{\mathbf{X}}(\mathbf{X}) = 1$. Up to a constant multiple we have

$$\omega = \frac{dx \wedge dy \wedge d\theta}{y^2},$$

where we have written $\mathbf{x} = (z, \xi)$ and $z = x + iy$, $\xi = ye^{i\theta}$ (the factor y is there to make it a unit vector): this measure is easily confirmed to be invariant under both left and right action of $\mathrm{PSL}_2\,\mathbb{R}$ on $\mathbf{H}$.

Occasionally, we will need a metric and not just a measure on $\mathrm{PSL}_2\,\mathbb{R}$; we will use the metric that corresponds under Φ to the Riemannian structure

$$\frac{dx^2 + dy^2}{y^2} + d\theta^2.$$

This metric is invariant under left action of $\mathrm{PSL}_2\,\mathbb{R}$ on $\mathbf{H}$, and as such does induce a metric on $\mathbf{X}$. It is *not invariant* under right action, and the flows u_+, u_- and g do not preserve lengths.

3 The Horocycle Flow is Ergodic

Theorem 1 *[Hed36] The positive and the negative horocycle flows are ergodic.*

We will show this for the positive ergodic flow. To prove Theorem 1 we will show that any $f \in L^2(\mathbf{X})$ invariant under the horocycle flow is constant almost everywhere. Indeed, if the positive horocycle flow is not ergodic then then there is a measurable set $\mathbf{Y} \in \mathbf{X}$ of positive but not full measure that is invariant under U_+ and the characteristic function $\mathbf{1_Y}$ provides a nonconstant invariant element of $L^2(\mathbf{X})$.

Lemma 2 *For $f \in L^2(\mathbf{X})$, $A \in \mathrm{PSL}_2\,\mathbb{R}$ and $\mathbf{x} \in \mathbf{X}$, let $(T_A f)(\mathbf{x}) := f(\mathbf{x} * A)$. Then the function $F_f : \mathrm{PSL}_2\,\mathbb{R} \to \mathbb{R}$ defined by*

$$F_f(A) = \int_{\mathbf{X}} f(\mathbf{x}) f(\mathbf{x} * A) \omega_{\mathbf{X}}(d\mathbf{x}) :=< f, T_A f >$$

is

(a) uniformly continuous and

(b) bi-invariant under U_+, i.e., invariant under the left and the right action of U_+ on $\mathrm{PSL}_2\,\mathbb{R}$.

Proof of Lemma 2 (a) Choose $\varepsilon > 0$. Since the continuous functions with compact support are dense in $L^2(\mathbf{X})$, we can find a function $g \in C_c(\mathbf{X})$

with $||f-g||_2 < \varepsilon/3$. The fact that such a g is uniformly continuous means that $\exists \delta > 0$ such that

$$d(A,B) < \delta \Rightarrow ||T_A g - T_B g||_2 < \frac{\varepsilon}{3}$$

So when $d(A,B) \leq \delta$ we have

$$\|T_A f - T_B f\|_2 \leq \|T_A f - T_A g\|_2 + \|T_A g - T_B g|_2 + \|T_B g - T_B f\|_2 \leq \varepsilon$$

We see that $A \mapsto T_A f$ is a uniformly continuous map $\mathrm{PSL}_2\,\mathbb{R} \to L^2(\mathbf{X})$ and (a) follows.

(b) Biinvariance reflects the invariance of Haar measure on $\mathrm{PSL}_2\,\mathbb{R}$ under left and right translation: for $A \in \mathrm{PSL}_2\,\mathbb{R}$ we have

$$\begin{aligned}
F_f(AU_+^t) &= \int_{\mathbf{X}} f(\mathbf{x}) f(\mathbf{x} * (AU_+^t)) \omega_{\mathbf{X}}(d\mathbf{x}) = \int_{\mathbf{X}} f(\mathbf{x}) f((\mathbf{x} * A) u_+(t)) \omega_{\mathbf{X}}(d\mathbf{x}) \\
&= \int_{\mathbf{X}} f(\mathbf{x}) f(\mathbf{x} * A) \omega_{\mathbf{X}}(d\mathbf{x}) = F_f(A); \\
F_f(U_+^t A) &= \int_{\mathbf{X}} f(\mathbf{x}) f(\mathbf{x} * (U_+^t A)) \omega_{\mathbf{X}}(d\mathbf{x}) = \int_{\mathbf{X}} f(\mathbf{x} * A^{-1}) f(\mathbf{x} * U_+^t) \omega_{\mathbf{X}}(d\mathbf{x}) \\
&= \int_{\mathbf{X}} f(\mathbf{x} * A^{-1}) f(\mathbf{x} u_+(t) \omega_{\mathbf{X}}(d\mathbf{x}) = \int_{\mathbf{X}} f(\mathbf{x} * A^{-1}) f(\mathbf{x}) \omega_{\mathbf{X}}(d\mathbf{x}) \\
&= \int_{\mathbf{X}} f(\mathbf{x}) f(\mathbf{x} * A) \omega_{\mathbf{X}}(d\mathbf{x}) = F_f(A)
\end{aligned}$$

Proof of Theorem 1. What do the biorbits of U_+ look like in $\mathrm{PSL}_2\,\mathbb{R}$? Using our identification $\Phi : \mathrm{PSL}_2\,\mathbb{R} \to \mathbf{H}$, we can think of the biorbits as living in $\mathbf{H}$ with geometry represented in figure 3. More specifically, there are two kinds of biorbits. The first (exceptional) kind of biorbit consists of all upwards pointing vertical vectors anchored at points $z = x + iy$ with a given y-coordinate. The union of these orbits forms a plane $\mathbf{V} \subset \mathbf{H}$. The other (generic) biorbits consist of the vectors defining horocycles of a given radius tangent to the x-axis: each such biorbit is diffeomorphic to a plane.

In particular, all the biorbits are closed, and there is nothing to prevent the existence of nonconstant continuous functions on $\mathbf{H}$ that are constant on biorbits. But our function F_f is *uniformly* continuous, and that changes the situation: every uniformly continuous function on $\mathbf{H}$ that is constant on biorbits *is* constant. What we need to see is that for every $\epsilon > 0$,

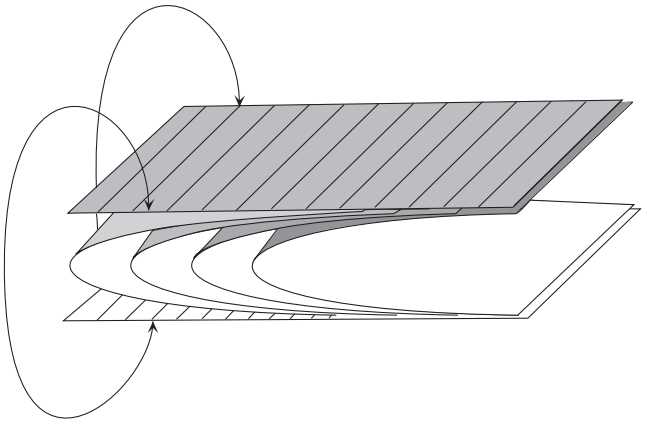

Figure 3: Since the left and the right action of U_+ commute, the biorbits are homeomorphic to $\mathbb{R}^2$, except the orbits on which the two actions coincide. Viewed in $\mathbf{H}$, the orbits are the 1-parameter family of folded planes (the ham slices in the sandwich). The top and the bottom planes should be identified; they represent the orbits formed of vertical upwards pointing vectors; those orbits are lines, as drawn in the planes. The salient features of the figure is that any two points of the top plane are within ϵ of a single biorbit (in fact all of those with folds sufficiently far to the left), and every biorbit comes arbitrarily close to a biorbit consisting of vertical upwards pointing vectors.

• any two elements of $\mathbf{V}$ can be approximated to within ϵ by a single 2-dimensional biorbit, and

• that any 2-dimensional biorbit is within ϵ of some element of $\mathbf{V}$.

These features are illustrated, but not proved, by figure 3. The proofs are the content of the two parts of figure 4:

Since F_f is uniformly continuous, for any $\epsilon > 0$ we can find a δ such that $d(\mathbf{z}, \mathbf{z}') < \delta \Rightarrow \|F_f(\mathbf{z}) - F_f(\mathbf{z}')\| < \epsilon/2$. Choose any two points $\mathbf{z} = (x + iy, \xi), \mathbf{z}' = (x' + iy', \xi') \in \mathbf{V}$, and assume without loss of generality that $y > y'$. Choose η a *nonvertical* vector for which $d((x+iy, \xi), (x+iy, \eta)) < \delta$. Set $\mathbf{w}'' = (x'' + iy', \eta')$ is the point on the the positive horocycle through $\mathbf{w} = (x + iy, \eta)$ at "height" y', and $w' = (x' + iy', \eta')$. The vector η' is *more* vertical than η, so $d((x'' + iy', \eta'), (x'' + iy', \xi')) < \delta$. Further, w' and w'''

Figure 4: Left: Any two upward-pointing vertical unit vectors ξ, ξ' can be approximated by elements of the same biorbit. Right: Any orbit contains vectors arbitrarily close to upward-pointing vertical vectors.

belong to the same biorbit. Thus

$$\begin{aligned}
&\|F_f(\mathbf{z}) - F_f(\mathbf{z}')\| \\
&\leq \|F_f(\mathbf{z}) - F_f(\mathbf{w})\| + \|F_f(\mathbf{w}) - F_f(\mathbf{w}'')\| + \|F_f(\mathbf{w}'') - F_f(\mathbf{w}')\| \quad + \|F_f(\mathbf{w}') - F_f(\mathbf{z}')\| \\
&\leq \frac{\epsilon}{2} + 0 + 0 + \frac{\epsilon}{2} = \epsilon.
\end{aligned}$$

The right side of fig **??** shows that every two-dimensional biorbit contains vectors arbitrarily close to vertical, ie. arbitrarily close to $\mathbf{V}$ just taking the vector perpendicular to the horocycle sufficiently close to the real axis.

Using biinvariance, we see that F_f is constant on $\mathbf{V}$; evidently this constant is $\|f\|_2^2 = F_f(\mathbf{x}_0)$. By the Cauchy-Schwarz inequality (and the fact that we chose f real), we see that for all $A \in \mathrm{PSL}_2(\mathbb{R})$, $T_A f = \pm f$ in $L^2(\mathbf{X})$ with the sign depending continuously on A. Since $\mathrm{PSL}_2(\mathbb{R})$ is connected, f is a constant element of $L^2(\mathbf{X})$ and the theorem follows. ■

4 Equidistribution of the horocycle flow

The main result of this paper is theorem 3.

Theorem 3 *[DS84] Let X be a complete hyperbolic surface of finite area. Then every horocycle on $\mathbf{X}$ is either periodic or equidistributed in $\mathbf{X}$.*

This theorem is evidently a much stronger statement than that the horocycle flow is ergodic, or even that it is uniquely ergodic. It is not an "almost everywhere"statement, but rather it asserts that *every horocycle* is either periodic or equidistributed in $\mathbf{X}$.

Note that this depends crucially on the fact that horocycles have geodesic curvature 1. The statement is false for geodesics (geodesic curvature 0): geodesics can do all sorts of things other than being periodic or equidistributed. For instance, they can spiral towards a closed geodesic, or be dense in a geodesic lamination, or spiral towards a geodesic lamination. Curves with constant geodesic curvature < 1 stay a bounded distance away from a geodesic, and hence can do more or less the same things as geodesics; in particular, they do not have to be periodic or equidistributed.

On the other hand, curves with geodesic curvature > 1 are always periodic, hence never equidistributed.

The horocycle flow is also distinguished from flows along curves of geodesic curvature less than 1 by having entropy 0. We will not define entropy here, but whatever definition you use it is clear that if you speed up a flow by a factor $\alpha > 0$, the entropy will be multiplied by α. But the formula

$$G(-s)U^+(t)G(s) = U^+(\exp(-s)t)$$

shows that the horocycle flow is conjugate to itself speeded up by $\exp(-s)$, thus its entropy must be 0.

5 The Geometry of Flows in H

The geodesic flow in $\mathbf{X}$ has *stable* and *unstable* foliations: two points $\mathbf{x}_1, \mathbf{x}_2 \in \mathbf{X}$ belong to the same leaf of the stable foliation if $d(\mathbf{x}_1 g(t), \mathbf{x}_2 g(t))$ is bounded as $t \to +\infty$, and they belong to the same leaf of the unstable foliation if $d(\mathbf{x}_1 g(t), \mathbf{x}_2 g(t))$ is bounded as $t \to -\infty$. These foliations are very easy to visualize in $T^1\mathbf{H}$, as shown in Figure 5.

Note that the stable leaves are fixed by the positive horocycle flow: the positive horocycles are the curves orthogonal to the geodesics in a leaf. Similarly, the unstable manifolds are fixed by the negative geodesic flow, and the negative horocycles in a leaf are the curves orthogonal to the geodesics in that leaf.

On the other hand, the positive horocycles are transverse to the unstable manifolds, and positive horocycle flow does not send unstable leaves to unstable leaves.

Let $S_{\mathbf{x}}$ be the unstable manifold of the geodesic flow through $\mathbf{x}$. Define

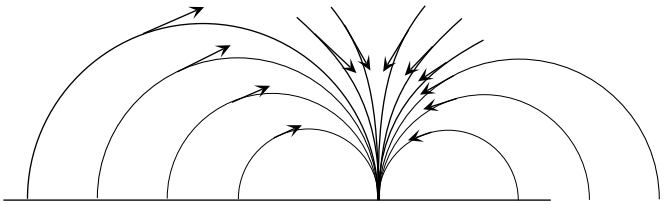

Figure 5: The tangent vectors to geodesics ending at the same point at infinity form one leaf of the stable foliation for the geodesic flow. Similarly, the tangent vectors to geodesics emanating from a point at infinity form a leaf of the unstable foliation. In $\mathbf{X}$, these leaves are tangled up in some very complicated way (after all, most geodesics are dense, never mind their stable and unstable manifolds). But clearly each leaf is an immersed smooth surface, hence of measure 0 in $\mathbf{X}$.

$S_{\mathbf{x}}(a,b) \subset S_{\mathbf{x}}$ by

$$S_{\mathbf{x}}(a,b) = \{\mathbf{x}u_-(r)g(s), |r| \leq a, |s| \leq b\}.$$

We will refer to $S_{\mathbf{x}}(a,b)$ as a "rectangle"; it isn't really: it is a quadrilateral bounded by two arcs of geodesic of length $2b$, and be two arcs of negative horocycle, of length respectively ae^b and ae^{-b} (see figure 6).

Further we define the "box" $W_{\mathbf{x}}(a,b,c) \subset \mathbf{X}$ as the region obtained by flowing along positive horocycles from $S_{\mathbf{x}}(a,b)$ until you hit $S_{\mathbf{x}u_+(c)}$. Because $S_{\mathbf{x}u_+(c)}$ is actually dense in $\mathbf{X}$, you have to understand the flow as taking place in the universal covering space $\mathbf{H}$, and then projecting the "box" to $\mathbf{X}$ (see figure 6 again).

Each surface $S_{\mathbf{x}}$ is invariant under geodesic flow, but the "rectangles" $S_{\mathbf{x}}(a,b)$ are not; instead we have

$$S_{\mathbf{x}}(a,b)g(s) = S_{\mathbf{x}g(s)}(e^s a, b).$$

Moreover, the surfaces $S_{\mathbf{x}u_+(s)}$ foliate a neighborhood of the positive horocycle $\mathbf{x} * U_+$ through $\mathbf{x}$, and thus there is a function $\alpha_{\mathbf{x}}(\mathbf{y},t) : \mathbb{R} \to \mathbb{R}$ for $\mathbf{y} \in S_{\mathbf{x}}$ (for its precise domain, see below) such that

$$\mathbf{y}u_+(\alpha_{\mathbf{x}}(\mathbf{y},t)) \in S_{\mathbf{x}u_+(t)}.$$

as sketched in Figure 6.

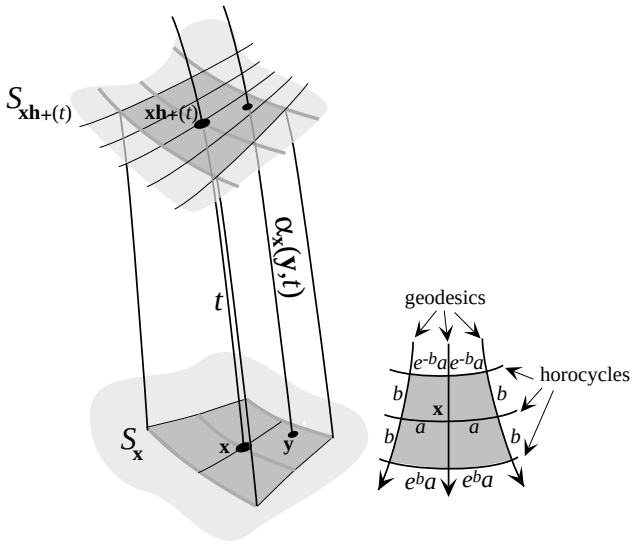

Figure 6: The surfaces $S_{\mathbf{x}u_+(s)}$ foliate a neighborhood of the positive horocyclic orbit of $\mathbf{x}$. Thus, for every t and every $\mathbf{y} \in S_{\mathbf{x}}$ sufficiently close to $\mathbf{x}$, there is a time $\alpha_{\mathbf{x}}(\mathbf{y}, t)$ such that the positive horocycle $\mathbf{y}u_+(\mathbb{R})$ intersects $S_{\mathbf{x}u_+(t)}$.

The function $t \mapsto \alpha_{\mathbf{x}}(\mathbf{y}, t)$ is defined in $[0, T(\mathbf{y}))$ for some $T(\mathbf{y})$ that tends to ∞ as $\mathbf{y} \to \mathbf{x}$. Moreover, the function is C^∞ (actually real-analytic) by the implicit function theorem, and $\frac{d}{dt}\alpha_{\mathbf{x}}(\mathbf{y}, t)$ tends to 1 as $\mathbf{y} \to \mathbf{x}$, so that

$$\lim_{\mathbf{y}\to\mathbf{x}} \frac{\alpha_{\mathbf{x}}(\mathbf{y}, t)}{t} \to 1.$$

It isn't often that you can replace the implicit function theorem by an explicit formula, but this does occur here.

Lemma 4 *If* $\mathbf{y} = \mathbf{x}u_-(r)g(t)$, *then*

$$\alpha_{\mathbf{x}}(\mathbf{y}, s) = \frac{s}{e^t(1 - rs)}. \tag{2}$$

In particular, $\alpha_{\mathbf{x}}(\mathbf{y}, s)$ *is defined in* $\mathbf{y} \in S_{\mathbf{x}}(a, b)$ *for all* $s < 1/a$ *and all* b, *and*

$$\lim_{\mathbf{y}\to\mathbf{x}} \frac{d}{ds}\alpha_{\mathbf{x}}(\mathbf{y}, s) = 1, \quad \lim_{\mathbf{y}\to\mathbf{x}} \frac{\alpha_{\mathbf{x}}(\mathbf{y}, s)}{s} = 1.$$

Proof. This is a matter of solving the equation

$$\mathbf{x}u_-(r)g(t)u_+(\alpha_{\mathbf{x}}(y,s)) = \mathbf{x}u_+(s)u_-(\rho)g(\tau),$$

i.e., the matrix equation

$$\begin{bmatrix}1 & 0\\ r & 1\end{bmatrix}\begin{bmatrix}e^{t/2} & 0\\ 0 & e^{-t/2}\end{bmatrix}\begin{bmatrix}1 & \alpha\\ 0 & 1\end{bmatrix} = \begin{bmatrix}1 & s\\ 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\ \rho & 1\end{bmatrix}\begin{bmatrix}e^{\tau/2} & 0\\ 0 & e^{-\tau/2}\end{bmatrix}.$$

This is a system of 3 equations (because the determinants are all 1) in 3 unknowns ρ, τ and α. Just multiply out and check. ■

The central result here is the following:

Lemma 5 *There exists a constant C such that for all $0 < \delta < 1/2$, all $t > 0$, all $\mathbf{y} \in S_{\mathbf{x}}(\frac{\delta}{t}, \delta)$ and all $0 \le s \le t$ we have*

$$d(\mathbf{x}u_+(s), \mathbf{y}u_+(\alpha_{\mathbf{x}}(\mathbf{y}, s)) \le C\delta.$$

Note that $\alpha_{\mathbf{x}}(\mathbf{y}, s)$ is defined for $s \le t$ when $\mathbf{y} \in S_{\mathbf{x}}(\frac{\delta}{t}, \delta)$ and $\delta \le 1/2$, since for the factor $1 - rs$ from the denominator of formula 2, we have $r \le \delta/t$ and $s \le t$, so $1 - rs \ge 1 - \delta^2 = 3/4$.

Proof. The proof essentially consists of gazing at Figure 7. Almost everything in that figure comes from the fact that the geodesic flow takes horocycles to horocycles; moreover, geodesic flow for time t maps a segment of positive horocycle of length l to one of length $e^{-t}l$, and a segment of negative horocycle of length l to one of length $e^t l$.

Two points $\mathbf{x}, \mathbf{y}$ with $\mathbf{y} \in S_{\mathbf{x}}(\delta/t, \delta)$ flow under the geodesic flow for time $\log t$ to two points $\mathbf{x}' = \mathbf{x}g(\log t)$ and $\mathbf{y}' = \mathbf{y}g(\log t)$; note that $\mathbf{y}' \in S_{\mathbf{x}'}(\delta, \delta)$, so certainly $d(x', y') \le 2\delta$. Then under positive horocycle flow (for different times) these points flow to points

$$\mathbf{x}'' = \mathbf{x}'u_+(s\delta) \quad \text{and} \quad \mathbf{y}'' \in S_{\mathbf{x}''}.$$

By the argument in the caption of figure 7, there exists a universal constant C such that $d(\mathbf{x}'', \mathbf{y}'') \le 2Cd(\mathbf{x}, \mathbf{y})$. Finally, use the geodesic flow back, i.e., for time $-\log t$, to find points

$$\mathbf{x}''' = \mathbf{x}u_+(s), \ \mathbf{y}''' = \mathbf{y}u_+(\alpha_{\mathbf{x}}(\mathbf{y}, s)).$$

Since backwards geodesic flow in a single unstable manifold is contracting, we find $d(x''', y''') \le 2C\delta$.

■

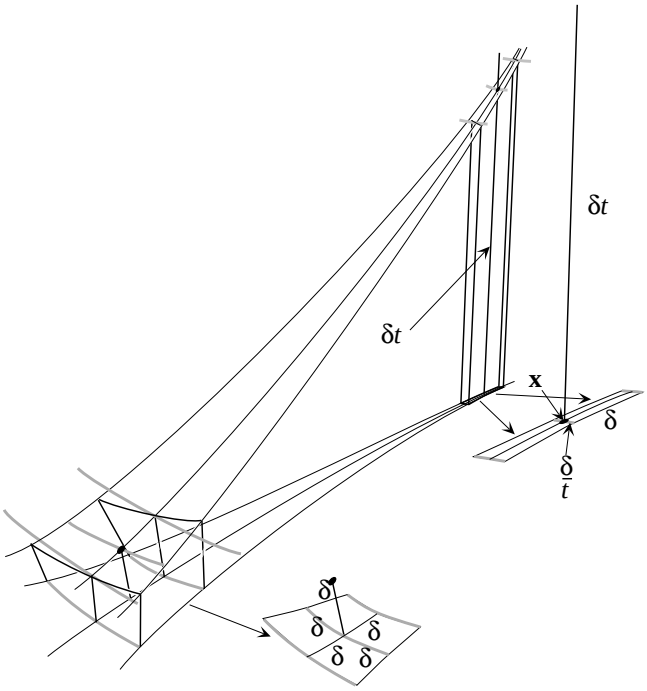

Figure 7: The skinny "rectangle" $S_{\mathbf{x}}(\frac{\delta}{t}, \delta)$ becomes under the geodesic flow for time $\log t$ the "square" $S_{\mathbf{x}g(\log t)}(\delta, \delta)$, and the box $W_{\mathbf{x}}(\frac{\delta}{t}, \delta, \delta t)$ becomes the box $W_{\mathbf{x}g(\log t)}(\delta, \delta, \delta)$. The geometry of $W_{\mathbf{x}g(\log t)}(\delta, \delta, \delta)$ is standard: it depends only on δ. In particular, the positive horocyclic flow from the bottom $S_{\mathbf{x}}(\frac{\delta}{t}, \delta)$ of the box is defined since $\delta \leq 1/2$, hence C^∞, hence Lipschitz with a universal constant C.

6 Geometry of hyperbolic surfaces and cusps

Let X be a complete hyperbolic surface. If such a surface is not compact, it has finitely many *cusps*. Every cusp c is surrounded by closed horocycles, and the open region bounded by the horocycle of length 2 is a neighborhood N_c isometric to the region $2\mathbb{Z}\backslash\{y \geq 1\}$ that is embedded in X, moreover, if c, c' are distinct cusps, then $N_c \cap N_{c'} = \emptyset$. If $\mathbf{X}$ has finite area and since each of the disjoint neighborhoods N_c of the cusps has area 2, there are only finitely many cusps. [Hub06].

We know everything about the standard cusp $\{y \geq 1\}/2\mathbb{Z}$, in particular that any geodesic that enters it will leave it again unless it goes directly to the cusp, i.e., unless it is a vertical line. Thus the same holds for all N_c.

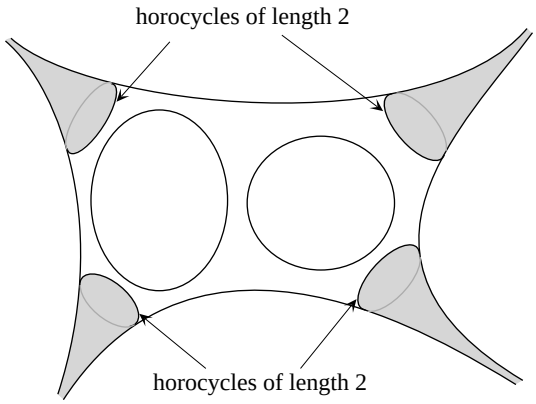

Figure 8: A non-compact complete hyperbolic surface is always non-compact in the same way: it has cusps c with disjoint standard neighborhoods N_c isometric to $\{y \geq 1\}/2\mathbb{Z}$, hence bounded by horocycles of length 2. Note that the only way a geodesic $\gamma(t)$ can stay in such a neighborhood for all $t \geq t_0$ is to head straight to the cusp. Each cusp has a stable manifold in X, and the geodesics that do not return infinitely many times to $X_c = X - \cup_c N_c$ are those that belong to one of these stable manifolds.

If X has finite area, then the complement of these neighborhoods is compact:

$$X_c = X - \bigsqcup_{\text{cusps } c \text{ of } X} N_c$$

is a compact set. Denote by $\mathbf{X}_c$ the corresponding part of $\mathbf{X}$. The injectivity radius is bounded below on $\mathbf{X}_c$, so there is a number $\delta_{\mathbf{X}} > 0$ such that for every $\mathbf{x} \in \mathbf{X}_c$ the box $W_{\mathbf{x}}(\delta_{\mathbf{X}}, \delta_{\mathbf{X}}, \delta_{\mathbf{X}})$ is embedded in $\mathbf{X}$.

Now, suppose that $\mathbf{X}$ has finite measure. Then if $(x, \xi) \in \mathbf{X}$ is a point through which the positive horocycle is not periodic, the geodesic through (x, ξ) does not go forward to a cusp and hence must enter $\mathbf{X}_c$ infinitely many times.

All periodic horocycles are homotopic to horocycles surrounding cusps. Indeed, if we apply geodesic flow to a positive horocycle, it will become arbitrarily short, thus will either be contained in a neighborhood of a cusp or in a contractible subset of $\mathbf{X}$. No horocycle in such a contractible subset is closed, thus the horocycle is homotopic to a horocycle surrounding a cusp.

Thus, the set $\mathbf{P}_{\mathbf{X}} \subset \mathbf{X}$ of points defining periodic positive horocycles is

the union of the stable manifolds of the cusps (for the geodesic flow).

Lemma 6 *The set* $\mathbf{P_X}$ *has measure zero in* $\mathbf{X}$ *for the measure* $\omega_{\mathbf{X}}$.

Proof. There are finitely many cusps, and each has a stable manifold which is a smooth immersed surface, certainly of 3-dimensional measure 0. ■

Although we have used the fact that $\mathbf{X}$ has finite area, the result is true for every complete hyperbolic surface, since such a surface can have only countably many cusps.

7 A sequence of good times

In this section we prove a result, still a bit weaker than theorem 3, though it does prove theorem 3 when X is compact.

Theorem 7 *Let* X *be a complete hyperbolic surface of finite area,* $\mathbf{X}$ *be its unit tangent bundle. For all* $\mathbf{x} \notin \mathbf{P_X}$*, there then exists a sequence* $T_n \to \infty$ *such that for any function* $f \in C_c(\mathbf{X})$ *we have*

$$\lim_{n\to\infty} \frac{1}{T_n} \int_0^{T_n} f(\mathbf{x}u_+(t))dt = \int_{\mathbf{X}} f d\omega_{\mathbf{X}}.$$

The proof will take the remainder of this section.

Proof. Let T_n be any increasing sequence tending to ∞ such that $\mathbf{x}g(t) \in \mathbf{X}_c$. Such a sequence exists because $\mathbf{x} \notin \mathbf{P_X}$.

Choose $\epsilon > 0$, and $f \in C_c(\mathbf{X})$; without loss of generality we may assume $\sup |f| = 1$ and that $\epsilon < 1$.

We have already defined $\delta_{\mathbf{X}}$. We need two more $\delta' s$, to be specified in lemmas 8 and 9.

Lemma 8 *There exists* $\delta_f > 0$ *such that for all* $t > 0$*, if* $\mathbf{z} \in S_{\mathbf{x}}(\delta_f/t, \delta_f)$ *and* $0 \le s \le t$*, then*

$$|f(\mathbf{x}u_+(s)) - f(\mathbf{z}u_+(\alpha_{\mathbf{x}}(\mathbf{z}, s)))| < \epsilon.$$

Proof. This follows immediately from proposition 5 and the uniform continuity of f. ■

Lemma 9 *There exists δ_α such that for all $t > 0$, if $\mathbf{z} \in S_{\mathbf{x}}(\delta_\alpha/t, \delta\alpha)$ and $0 \le s \le t$, then*

$$|\alpha'_{\mathbf{x}}(\mathbf{z}, s) - 1| < \epsilon.$$

Proof. One could derive this from the implicit function theorem, but we might as well use our explicit formula (2) for α. For $\mathbf{z} = \mathbf{x}u_-(r)g(u) \in S_{\mathbf{x}}(\delta_\alpha/t, \delta_\alpha)$ we have

$$\alpha'_{\mathbf{x}}(\mathbf{y}, s) = \frac{1}{e^u(1-rs)^2},$$

and since $|r| \le \delta_\alpha/t$, $|u| \le \delta_\alpha$ and $s \le t$,

$$\frac{1}{e^\delta_\alpha(1+\delta_\alpha)^2} \le \alpha'_{\mathbf{x}}(\mathbf{y}, s) \le \frac{e^\delta_\alpha}{(1-\delta_\alpha)^2}.$$

Clearly we can choose δ_α so that

$$\left|\frac{1}{e^{\delta_\alpha}(1+\delta_\alpha)^2} - 1\right| < \epsilon, \quad \left|\frac{e^{\delta_\alpha}}{(1-\delta_\alpha)^2} - 1\right| < \epsilon.$$

■

Set $\delta = \inf(\delta_{\mathbf{X}}, \delta_f, \delta_\alpha)$, and $\eta = \omega_{\mathbf{X}}(W_{\mathbf{x}}(\delta, \delta, \epsilon\delta))$.

Proposition 10 *There exists a $\tilde{T}$ and a set $\mathbf{Y} \subset \mathbf{X}$ with $\omega_{\mathbf{X}}(\mathbf{Y}) > 1 - \eta$ such that for all $T \ge \tilde{T}$ and all $\mathbf{y} \in \mathbf{Y}$ we have*

$$\left|\frac{1}{T}\int_0^T f(\mathbf{y}u_+(t))dt - \int_{\mathbf{X}} f d\omega_{\mathbf{X}}\right| < \epsilon.$$

Proof. By the ergodic theorem, the family of functions

$$g_T(\mathbf{y}) = \frac{1}{T}\int_0^T f(\mathbf{y}u_+(t))dt$$

converges almost everywhere as $T \to \infty$, and since horocycle flow is ergodic, it converges almost everywhere to $\int_{\mathbf{X}} f d\omega_{\mathbf{X}}$ (this is where we use Theorem 1). By Egorov's theorem, there exists a set $\mathbf{Y}$ of measure at least $1 - \eta$ such that the g_T converge uniformly on $\mathbf{Y}$; omitting a set of measure 0 from $\mathbf{Y}$, the family g_T converges uniformly to $\int_{\mathbf{X}} f d\omega_{\mathbf{X}}$ on Y. This is the assertion of Proposition 10. ■

We can now choose

1. an n_0 such that $T_n > \tilde{T}$ for $n > n_o$. Then for $n > n_0$ we can select:

2. a sequence $\mathbf{y}_n \in \mathbf{Y} \cap W_{\mathbf{x}}(\delta/T_n, \delta, \epsilon\delta T_n)$. Indeed, we have

$$\omega_{\mathbf{X}}(W_{\mathbf{x}}(\delta/T_n, \delta, \epsilon\delta T_n)) = \eta,$$

since it is the inverse image of $W_{\mathbf{x}u_+(T_n)}(\delta, \delta, \epsilon\delta)$ by the geodesic flow at time T_n. We have

$$\omega_{\mathbf{X}}(W_{\mathbf{x}u_+(T_n)}(\delta, \delta, \epsilon\delta)) = \eta$$

since $\mathbf{x}u_+(T_n) \in \mathbf{X}_c$ and $\delta \leq \delta_{\mathbf{X}}$. Geodesic flow for a fixed time is a measure-preserving diffeomorphism, so $W_{\mathbf{x}}(\delta/T_n, \delta, \epsilon\delta T_n)$ must intersect $\mathbf{Y}$ which has volume $> 1 - \eta$.

3. sequences $\mathbf{z}_n \in S_{\mathbf{x}}(\delta/T_n, \delta)$ and $\epsilon'_n \leq \epsilon$ such that $\mathbf{z}_n u_+(\epsilon'_n \delta T_n) = \mathbf{y}_n$. This is just what it means to say $\mathbf{y}_n \in W_{\mathbf{x}}(\delta/T_n, \delta, \epsilon\delta T_n)$.

The organizing principle is now to write for $n > n_0$

$$\begin{aligned}
&\left|\frac{1}{T_n}\int_0^{T_n} f(\mathbf{x}u_+(t))dt - \int_{\mathbf{X}} f(\mathbf{w})\omega_X(d\mathbf{w})\right| \leq \\
&\quad\left|\frac{1}{T_n}\int_0^{T_n} f(\mathbf{x}u_+(t))dt - \frac{1}{T_n}\int_0^{T_n} f\Big(\mathbf{z}_n u_+\big(\alpha_{\mathbf{x}}(\mathbf{z}_n, t)\big)\Big)\right| + \\
&\quad\left|\frac{1}{T_n}\int_0^{T_n} f\Big(\mathbf{z}_n u_+\big(\alpha_{\mathbf{x}}(\mathbf{z}_n, t)\big)\Big) - \frac{1}{T_n}\int_0^{T_n} f\Big(\mathbf{z}_n u_+\big(\alpha_{\mathbf{x}}(\mathbf{z}_n, t)\big)\Big)\alpha'_{\mathbf{x}}(\mathbf{z}_n, t)dt\right| + \\
&\quad\left|\frac{1}{T_n}\int_0^{\alpha_{\mathbf{x}}(\mathbf{z}_n, T_n)} f(\mathbf{z}_n u_+(s))ds - \frac{1}{T_n}\int_0^{T_n} f(\mathbf{z}_n u_+(s))ds\right| + \\
&\quad\left|\frac{1}{T_n}\int_0^{T_n} f(\mathbf{z}_n u_+(s))ds - \frac{1}{T_n}\int_0^{T_n} f(\mathbf{y}_n u_+(s))ds\right| + \\
&\quad\left|\frac{1}{T_n}\int_0^{T_n} f(\mathbf{y}_n u_+(s))ds - \int_{\mathbf{X}} f(\mathbf{w})\omega_{\mathbf{X}}(d\mathbf{w})\right|.
\end{aligned}$$

To get from the second summand on the right to the third, we use the change of variables formula, setting $s = \alpha_{\mathbf{x}}(\mathbf{z}_n, t)$):

$$\int_0^{T_n} f\Big(\mathbf{z}_n u_+\big(\alpha_{\mathbf{x}}(\mathbf{z}_n, t)\big)\Big)\alpha'_{\mathbf{x}}(\mathbf{z}_n, t)dt = \int_0^{\alpha_{\mathbf{x}}(\mathbf{z}_n, T_n)} f(\mathbf{z}_n u_+(s))ds.$$

Each of the five terms above needs to be bounded in terms of ϵ.

1. Since $\delta < \delta_f$, we have $|f(\mathbf{x}u_+(t)) - f(\mathbf{z}_n u_+(\alpha_{\mathbf{x}}(\mathbf{z}_n, t)))| < \epsilon$, so

$$\left| \frac{1}{T_n} \int_0^{T_n} f(\mathbf{x}u_+(s))ds - \frac{1}{T_n} \int_0^{T_n} f(\mathbf{z}_n u_+(\alpha_{\mathbf{x}}(\mathbf{z}_n, t)))dt \right| < \epsilon.$$

2. Since $\delta < \delta_\alpha$, we have

$$\left| \frac{1}{T_n} \int_0^{T_n} f(\mathbf{z}_n u_+(\alpha_{\mathbf{x}}(\mathbf{z}_n, t)))dt - \frac{1}{T_n} \int_0^{T_n} f(\mathbf{z}_n u_+(\alpha_{\mathbf{x}}(\mathbf{z}_n, t)))\alpha'_{\mathbf{x}}(\mathbf{z}_n, t)dt \right|$$
$$\leq \frac{1}{T_n} \int_0^{T_n} \sup|f| \; |1 - \alpha'_{\mathbf{x}}(\mathbf{z}_n, t)|dt < \epsilon.$$

3. From $\delta < \delta_\alpha$, so $|\alpha' - 1| < \epsilon$, we get that $(1-\epsilon)T_n < \alpha_{\mathbf{x}}(\mathbf{z}_n, T_n) < (1+\epsilon)T_n$ and hence

$$\left| \frac{1}{T_n} \int_0^{\alpha_{\mathbf{x}}(\mathbf{z}_n, T_n)} f(\mathbf{z}_n u_+(s))ds - \frac{1}{T_n} \int_0^{T_n} f(\mathbf{z}_n u_+(s))ds \right| < \epsilon.$$

4. The points $\mathbf{z}_n$ and $\mathbf{y}_n$ are on the same positive horocycle, a distance $\epsilon'_n T_n$ apart for some $\epsilon'_n \leq \epsilon$. This leads to

$$\left| \frac{1}{T_n} \int_0^{T_n} f(\mathbf{z}_n u_+(s))ds - \frac{1}{T_n} \int_0^{T_n} f(\mathbf{y}_n u_+(s))ds \right|$$
$$= \left| \frac{1}{T_n} \int_0^{T_n} f(\mathbf{z}_n u_+(s))ds - \frac{1}{T_n} \int_{\epsilon'_n T_n}^{(1+\epsilon'_n)T_n} f(\mathbf{z}_n u_+(s))ds \right| \leq \frac{2\epsilon T_n}{T_n} = 2\epsilon.$$

5. Since $\mathbf{y}_n \in \mathbf{Y}$ and $T_n > \tilde{T}$, we have

$$\left| \frac{1}{T_n} \int_0^{T_n} f(\mathbf{y}_n u_+(s))ds - \int_{\mathbf{X}} f(\mathbf{w})\omega_{\mathbf{X}}(d\mathbf{w}) \right| < \epsilon.$$

■ This ends the proof of theorem 7. □

8 Proving equidistribution

The T_n are chosen to be a sequence of times tending to infinity such that $\mathbf{x}g(T_n) \in \mathbf{X}_c$. Thus if $\mathbf{X}$ is compact, the sequence T_n is an arbitrary sequence tending to infinity, and so equidistribution is proved in that case. Moreover, clearly theorem 7 shows that all non-periodic horocycles are dense in $\mathbf{X}$. But it doesn't quite prove that they are equidistributed when $\mathbf{X}$ is not compact; perhaps a horocycle could spend an undue amount of time near some cusp, and we could choose a different sequence of times T'_n also tending to infinity which would emphasize the values of f near that cusp. In fact, we will see in Section 9 that something like that does happen for random walks on horocycles.

We will now show that this does not happen for the horocycle flow itself.

Proposition 11 *Let ν be a probability measure on $\mathbf{X}$ invariant under the positive horocycle flow and such that $\nu(\mathbf{P}_\mathbf{X}) = 0$. Then $\nu = \omega_\mathbf{X}$.*

Proof. Without loss of generality, we can assume that ν is ergodic for the positive horocycle flow since any invariant probability measure is a direct integral of ergodic invariant probability measures ν_α with $\nu_\alpha(\mathbf{P}_\mathbf{X}) = 0$, so uniqueness for such invariant ergodic measures implies uniqueness for all such invariant measures. Choose $f \in C_c(\mathbf{X})$, and let $\mathbf{x} \in \mathbf{X}$ be a typical point for ν, i.e., a point of $\mathbf{X} - \mathbf{P}_\mathbf{X}$ such that

$$\int_\mathbf{X} f d\nu = \lim_{t\to\infty} \frac{1}{t} \int_0^t f(\mathbf{x}u_+(s))ds.$$

By the ergodic theorem, this is true of ν-almost every point, so such points $\mathbf{x}$ certainly exist. Such a point is one for which the horocycle flow is not periodic, so theorem 7 asserts that there exists a sequence $t_n \to \infty$ such that

$$\int_\mathbf{X} f d\nu = \lim_{t\to\infty} \frac{1}{t} \int_0^t f(\mathbf{x}u_+(s))ds = \lim_{n\to\infty} \frac{1}{t_n} \int_0^{t_n} f(\mathbf{x}u_+(s))ds = \int_\mathbf{X} f d\omega_\mathbf{X}.$$

Since this equality is true for every $f \in C_c(\mathbf{X})$, we have $\nu = \omega_\mathbf{X}$. ■

Now suppose that for some $\mathbf{x} \in \mathbf{X} - \mathbf{P}_\mathbf{X}$ and some $f \in C_c(\mathbf{X})$, we do not have

$$\int_\mathbf{X} f d\omega_\mathbf{X} = \lim_{t\to\infty} \frac{1}{t} \int_0^t f(\mathbf{x}u_+(s))ds.$$

We can consider the set of probability measures ν_t defined by

$$\int_{\mathbf{X}} f d\nu_t = \frac{1}{t}\int_0^t f(xu_+(s))ds.$$

On a non-compact space, the Riesz representation theorem says that set of Borel measures is the dual of the Banach space $C_0(X)$ of continuous functions vanishing at ∞ with the sup norm. The collection of probability measures ν_t is a subset of the unit ball, and the unit ball is compact for the weak topology. So if $\lim_{t\to\infty}\nu_t \neq \omega_{\mathbf{X}}$, there exists a measure $\nu \neq \omega_{\mathbf{X}}$ and a sequence $t_i \to \infty$ such that

$$\lim_{i\to\infty}\nu_{t_i} = \nu$$

in the weak topology.

Clearly ν is invariant under the horocycle flow and ergodic. So it might seem that $\omega_{\mathbf{X}} \neq \nu$ contradicts proposition 11. There is a difficulty with this argument when $\mathbf{X}$ is not compact. In that case the probability measures do not form a closed subset of the unit ball of $C_0(\mathbf{X})^*$; consider for instance the measures $\delta(x-n)$ on $\mathbb{R}$; as $n\to\infty$ they tend to 0 in the weak topology. Technically, the problem is that we can't evaluate measures on the continuous function 1, since this function doesn't vanish at infinity.

We need to show that ν is a probability measure. This follows from proposition 12 below. For $\rho \leq 2$, let $\mathbf{X}^\rho \subset \mathbf{X}$ be the compact part of $\mathbf{X}$ in which all periodic horocycles have length $\geq \rho$. So $\mathbf{X}_c = \mathbf{X}^2$

Proposition 12 . *For any $\epsilon > 0$, there exists $\rho > 0$ such that for all $\mathbf{x} \in \mathbf{X} - \mathbf{P}_{\mathbf{X}}$ we have*

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbb{1}_{\mathbf{X}^\rho}(\mathbf{x}u_+(s))ds > 1-\epsilon.$$

Proof. If c is a cusp of X, let N_c^ρ be the neighborhood of c bounded by the horocycle of length ρ. Recall from our discussion of the geometry of hyperbolic surfaces, that N_c^2 is isometric to a standard object: the part of $(2\mathbb{Z})\backslash\mathbf{H}$ where $y > 1$. Set γ to be the Moebius transformation $\gamma(z) = z/(z+1)$; the standard neighborhood is then isometric to the part of $\langle\gamma\rangle\backslash\mathbf{H}$ where $x^2 + (y-1)^2 \leq 1$. Moreover, any horocycle that doesn't tend to the cusp is equivalent by a change of variable commuting with γ to a horizontal

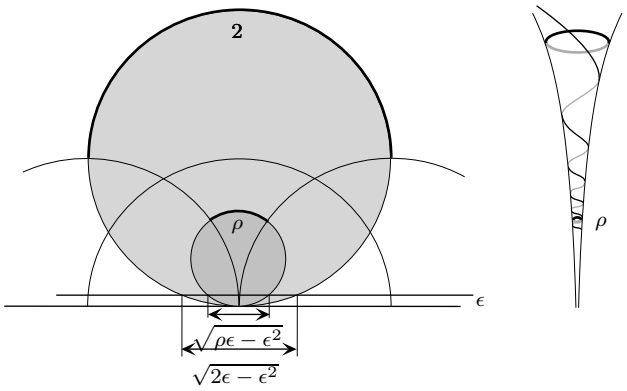

Figure 9: In H a neighborhhod of a cusp bounded by a horocycle corresponds to a disc tangent to the x-axis. In the figure on the left, we have represented the cusp by $\langle\gamma\rangle\backslash H$; without loss of generality we may set $\gamma(z) = z/(1+z)$. Then the disc of radius 1 centered at i corresponds to the neighborhood of the cusp bounded by the horocycle of length 2, and the disc of radius $\rho/2$ centered at $i\rho/2$ corresponds to the neighborhood bounded by a horocycle of length ρ. A horocycle that enters this neighborhood but does not go to the cusp can be, without loss of generality, represented by a line of equation $y = \epsilon$; it goes deeper and deeper into the cusp as $\epsilon \to 0$. The ratio of times spent in N^ρ to the time spent in $N^2 - N^\rho$ does not become large as the horocycle goes deeper in the cusp, but tends to a ratio depending only on ρ, which tends to 0 as ρ tends to 0. As horocycles go deeper and deeper in the cusp, they spiral more and more tightly in $N^2 - N^\rho$ and still spend approximately the same fraction of time in $N^2 - N^\rho$ as in N^ρ.

line. Of course lengths on such a horizontal line $y = \epsilon$ depend on ϵ, but ratios of lengths are the same as ratios of euclidean lengths.

A careful look at figure 9 shows that if a horocycle starts in $\mathbf{X}_c$, goes deep in the cusp, and comes out again, then the ratio of time spent in N^ρ to time spent in $N^2 - N^\rho$ is

$$\frac{\sqrt{\rho\epsilon - \epsilon^2}}{\sqrt{2\epsilon - \epsilon^2} - \sqrt{\rho\epsilon - \epsilon^2}} = \frac{\sqrt{\rho}}{\sqrt{2} - \sqrt{\rho}} + O(\epsilon). \tag{3}$$

Any non-periodic horocycle will eventually enter $\mathbf{X}_c$; by taking ρ sufficiently small, we can assure that afterwards it will spend a proportion of its time $< \epsilon$ outside of X^ρ. Proposition 12 follows. ■

Consider the measures

$$\nu_{\mathbf{x},T} = (f \mapsto \frac{1}{T}\int_0^T f(\mathbf{x}u_+(t)dt).$$

Proposition 13 *The accumulation set of* $\{\nu_{\mathbf{x},T}, T > 0\}$ *consists entirely of probability measures.*

Proof. Every accumulation point μ of the $\nu_{\mathbf{x},T}$ in $C_0(\mathbf{X})^*$ is a measure, and the only thing to show is that $\mu(\mathbf{X}) = 1$. Clearly $\mu(\mathbf{X}) \leq 1$, since for any $f \in C_0(X)$ and any $\mathbf{x}, T$ we have

$$\frac{1}{T}\int_0^T f(\mathbf{x}u_+(t))dt \leq \|f\|_\infty.$$

To see that $\mu(\mathbf{X}) \geq 1$, take $\epsilon > 0$ and ρ as in proposition 12. We can then find a function $f \in C_0(\mathbf{X})$ which coincides with $\mathbb{1}_{\mathbf{X}^\rho}$ on $\mathbf{X}^\rho$ and satisfies $0 \leq f \leq 1$ everywhere. Then

$$\mu(\mathbf{X}) = \sup_{g \in C_0(\mathbf{X})} \frac{\int_{\mathbf{X}} |g d\mu|}{\|g\|_\infty} \tag{4}$$

$$\geq \int_{\mathbf{X}} f d\mu \geq \liminf_{T\to\infty} \frac{1}{T}\int_0^T f(\mathbf{x}u_+(t))dt \tag{5}$$

$$\geq \liminf_{T\to\infty} \frac{1}{T}\int_0^T \mathbb{1}_{\mathbf{X}^\rho}(\mathbf{x}u_+(t))dt > 1 - \epsilon. \tag{6}$$

■

There is one last thing to check.

Proposition 14 *A measure* μ *in the limit set of* $\{\nu_{\mathbf{x},T}, T > 0\}$ *with* $\mathbf{x} \notin \mathbf{P}_\mathbf{X}$ *satisfies* $\mu(\mathbf{P}_\mathbf{X}) = 0$.

Proof. Suppose $\mu(\mathbf{P}_\mathbf{X}) > 0$, set $\epsilon = \mu(\mathbf{P}_\mathbf{X})/3$ and use proposition 12 to find a corresponding ρ. Find a compact subset $\mathbf{Q} \subset \mathbf{P}_\mathbf{X}$ with $\mu(\mathbf{Q}) > \frac{2}{3}\mu(\mathbf{P}_\mathbf{X})$, and find a time T such that

$$\mathbf{Q}g(T) \cap \mathbf{X}^\rho = \emptyset.$$

This is possible because $\mathbf{P}_\mathbf{X}$ consists of points in the stable manifolds of the cusps, so each point can be moved off $\mathbf{X}^\rho$, and since $\mathbf{Q}$ is compact it will leave $\mathbf{X}^\rho$ under the geodesic flow at some time T.

Let $\mathbf{U}$ be a neighborhood of $\mathbf{Q}$ such that $\mathbf{U}g(T) \cap \mathbf{X}^\rho = \emptyset$. For this neighborhood $\mathbf{U}$ of $\mathbf{Q}$, as for any neighborhood, there exists a sequence of times $T_n \to \infty$ such that

$$\frac{\lambda\{t \in [0, T_n] \mid \mathbf{x}u_+(t) \in \mathbf{U}\}}{T_n} > \frac{1}{2}\mu(\mathbf{Q}),$$

where λ is linear measure. Then the horocycle $t \mapsto \mathbf{x}g(T)u_+(t)$ must spend the same proportion of its time in $\mathbf{U}g(T)$, hence outside $\mathbf{X}^\rho$. But every non-periodic horocycle spends at least a proportion $1 - \mu(\mathbf{P}_X)/3$ in $\mathbf{X} - \mathbf{X}^\rho$, and this is a contradiction. ■

9 Horocycle flow on the modular surface

Let Γ be the 2-congruence subgroup of $\mathrm{SL}_2(\mathbb{R})$, so that $\mathbf{X} = \Gamma\backslash \mathrm{SL}_2\,\mathbb{R}$ is the unit tangent bundle over $X = \Gamma\backslash H$, which is the 3-times punctured sphere. Denote by $\pi_{\mathbf{X}} : \mathbf{H} \to \mathbf{X}$ and $\pi_X : \mathbf{X} \to X$ respectively the projections from $\mathbf{H} \cong \mathrm{PSL}(2, \mathbb{R})$ onto $\mathbf{X} \cong \Gamma\backslash \mathrm{PSL}(2, \mathbb{R})$ and from $\mathbf{X}$ onto X.

Lemma 15 *The hyperbolic surface X has area 2π, and the subset $X - X^\rho$ has area 3ρ.*

We leave the proof of this lemma to the reader.

It follows from lemma 15 that for every $\mathbf{x}_0 \notin \mathbf{P}_{\mathbf{X}}$ there exists for every sequence $\rho_n \to 0$, for every $\epsilon > 0$ and for every n sufficiently large, a time

$$T_n < (1 + \epsilon)\left(\frac{2\pi}{3\rho_n}\right)$$

such that $\pi_X(\mathbf{x}_0 u_+(T_n)) \in X - X^{\rho_n}$.

To use this result, we need to understand the region in H corresponding to X^ρ.

Lemma 16 *The inverse image in H of $X - X^\rho$ is the union of the horodisc $\mathrm{Im}\, z > 2/\rho$, and the union, for all rational numbers p/q of the discs of radius $\rho/(4q^2)$ tangent to the real axis at p/q.*

Choose $\alpha \in \mathbb{R} - \mathbf{Q}$, and consider the horocycle in $\mathbf{X}$ which is the image of the horocycle $\mathbf{z}_0 u_+(t)$ in $\mathbf{H}$, where $\mathbf{z}_0 = \alpha + 2i$, i.e., the image of the horocycle

represented by the circle of radius 1 tangent to the real axis at the irrational number α. A straightforward computation shows that

$$\mathbf{z}_0 u_+(T) = \left(\alpha - \frac{2T}{T^2+1}\right) + i\frac{2}{T^2+1}.$$

Set $\rho_n = 1/n$. This horocycle is not periodic, so it must enter $\mathbf{X} - \mathbf{X}^{\rho_n}$ at a sequence of times $T_n < (1+\epsilon)\left(\frac{2\pi n}{3}\right)$. Interpreting the cusps as rational numbers, this means that there exists an infinite sequence of rational numbers p_n/q_n and times $T_n < (1+\epsilon)\frac{2\pi n}{3}$ such that

$$\left|\alpha - \frac{p_n}{q_n}\right| = \frac{2T_n}{T_n^2+1} \leq \frac{T_n}{2nq_n^2} < (1+\epsilon)\frac{\pi n}{3nq_n^2} = (1+\epsilon)\frac{\pi}{3q_n^2}.$$

This is of course nothing to boast about. It has been known for over 100 years that for every irrational number α, there exist infinitely many coprime numbers numbers $p_n,\ q_n$ such that

$$\left|\alpha - \frac{p_n}{q_n}\right| < \frac{1}{\sqrt{5}q_n^2},$$

and that $1/\sqrt{5}$ is the smallest number for which this is true [Kin64]. Our analysis only gives the constant $\pi/3$, too large by a factor of more than 2.

One reason to take an interest in this result despite its weakness is that Ratner's theorem has many generalizations to situations where methods leading to the sharp results about diophantine approximations of irrational numbers are not available. In all settings, Ratner's theorem has "diophantine" consequences.

Clearly we cannot do better than improve the constant for all horocycles. But we can use the theory of diophantine approximations to improve the results above for almost every horocycle. In particular we can apply the following theorem.

Theorem 17 *[Kin64] If $g(x) : \mathbb{R}_+^* \to \mathbb{R}$ is a function such that $g(x)/x$ is increasing, then for almost every α, there exist infinitely many coprime integers p, q such that $|\alpha - \frac{p}{q}| < \frac{1}{qg(q)}$ if and only if the series*

$$\sum_{n=1}^{\infty} \frac{1}{g(n)}$$

diverges.

Let us see what this says about horocycles; we will specialize to the case where $g(n) = n\log(n)$. For almost every $\mathbf{x}_0$, the horocycle $\mathbf{x}_0 U_+$ lifts to a horocycle $\mathbf{z}_0 U_+$ in $\mathbf{H}$ tangent to the x-axis at an irrational number α belonging to the set of full measure from theorem 17. Changing the time parameterization by a constant, we may assume that $\mathbf{z}_0 = (\alpha + 2iR, -2iR)$ since this is in any case the worst point on the horocycle $\mathbf{z}_0 U_+$.

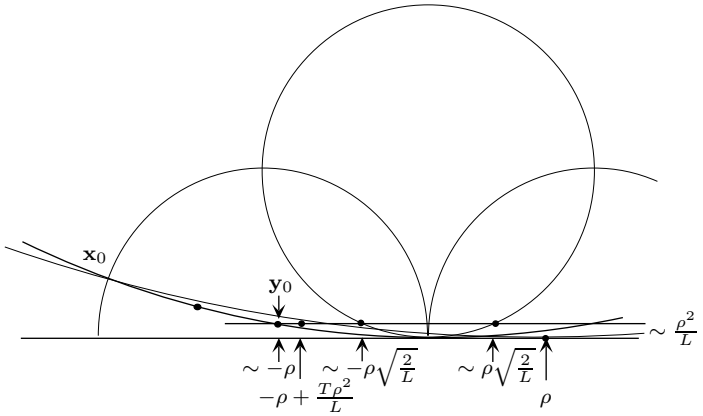

Figure 10: We can lift the horocycle $\mathbf{x}_0 U_+$ to $\mathbf{H}$; without loss of generality we may assume that the cusp c is at 0, and that the stabilizer of the cusp is generated by $z \mapsto z/(z+1)$. In that case, the horocycle of length 2 lifts to the circle of radius 1 centered at i, and the horocycle of length L lifts to the circle of radius $L/2$ centered at $Li/2$. We may take $x_0 = \pi_X(\mathbf{x}_0)$ to be anywhere on this horocycle; it will be convenient to place it at $\frac{-2L^2+4Li}{L^2+4}$. In that case, one fundamental domain on the horocycle goes from $\frac{-2L^2+4Li}{L^2+4}$ to $\frac{2L^2+4Li}{L^2+4}$. Our modified horocycle will join x_0 to the point $\rho > 0$ on the real axis. It is much easier to estimate lengths on this horocycle if we send ρ to infinity by a parabolic transformation that fixes 0, and hence all the horocycles tangent to the real axis at 0. If we perform this parabolic transformation, the point x_0 moves to a point y_0 on its horocycle which is approximately $-\rho + i\rho^2/L$, and the horocycle is a horizontal line, approximately the line $y = \rho^2/L$.

Let p_n/q_n be one of the good approximations to α guaranteed by theorem 17. Let ρ_n be the radius of the negative horocycle surrounding the cusp corresponding to p_n/q_n when the point $\mathbf{z}_0 u_+(T_n)$ is on the vertical line $x = p_n/q_n$. Further let us write

$$\mathbf{z}_0 u_+(T_n) = \xi_n + i\eta_n = 2R\left(\frac{T_n}{T_n^2+1} + \frac{i}{T_n^2+1}\right).$$

Then we have $T_n = \frac{\xi_n}{\eta_n}$ and $\eta_n = \frac{\rho_n}{2q_n^2}$ and by lemma **??** it follows that

$$T_n = \frac{\xi_n}{\eta_n} \leq \frac{1/(q_n^2 \log q_n)}{\rho_n/(2q_n^2)} = \frac{2}{\rho_n \log q_n}.$$

Moreoever $\eta_n = R - \sqrt{(}R^2 - \xi_n^2) \sim \frac{2\xi_n^2}{R}$ from which we can derive that $q_n \sim \frac{1}{2\log(\frac{1}{\rho_n})}$.

We have proved the following:

Theorem 18 *On the modular surface, for every $\epsilon > 0$ and for almost every horocycle $\mathbf{x}_o U_+$, there exists a sequence $\rho_n \to 0$ and times $T_n < (1+\epsilon)\frac{1}{\rho_n \log \frac{1}{\rho_n}}$ such that $\mathbf{x}_o u_+(T_n) \in \mathbf{X} - \mathbf{X}^{\rho_n}$.*

The theorem means that almost every nonperiodic horocycle enters $\mathbf{X} - \mathbf{X}^{\rho_n}$ much earlier than is implied for every nonperiodic horocycle by equidistribution.

This leads to a surprising result due to Breuillard [Bre05]: although nonperiodic horocycles are equidistributed, *any* uncentered random walk on the set of non-periodic horocycles almost surely is not.

Theorem 19 *Let μ be a probability measure on $\mathbb{R}$ with finite expectation and variance:*

$$0 \neq a = \int_{-\infty}^{\infty} t\mu(dt) < \infty \quad \textit{and} \quad b^2 = \int_{-\infty}^{\infty} (t-a)^2 \mu(dt) < \infty.$$

Denote by μ^{*m} the mth convolution of μ with itself. If $b > 0$, there exists a function $f \in C_c(\mathbf{X})$ with $\int_{\mathbf{X}} f(\mathbf{x})\omega_X(d\mathbf{x}) = 1$ such that for almost every $\mathbf{x}_0 \in \mathbf{X}$ we have

$$\liminf_{m\to\infty} \int_{-\infty}^{\infty} f(\mathbf{x}_0 u_+(t))\mu^{\star m}(dt) = 0.$$

Proof. Let $\alpha \in \mathbb{R}$ be an arbitrary element of the full Lebesgue measure set guaranteed by theorem 17. Suppose that the horocycle $\mathbf{z}_0 U_+$ through $\mathbf{z}_0 = (\alpha + 2iR, -2iR)$ in $\mathbf{H}$ projects to a horocycle in $\mathbf{X}$ containing $\mathbf{x}_0$. Now choose an approximating sequence of coprime $(p_n, q_n) \in \mathbb{Z}^2$ to α as guaranteed by theorem 17 above, and let T_n be the associated sequence of times. The measure $\mu^{\star m}$ is approximately the Gaussian of mean ma and standard deviation $\sqrt{m}b$. Let us choose an m such that:

1. $ma = T_n$ for one of the T_n given in theorem 18

2. the standard deviation of $\sigma(\mu^{\star m}) \sim b\sqrt{m}$ is much smaller than $1/\sqrt{\rho_n}$

This first condition can obviously be satisfied, and the second is also straightforward since

$$b\sqrt{m} \sim b\sqrt{\frac{T_n}{a}} \leq \frac{b}{\sqrt{a}}\sqrt{\frac{2}{\rho_n \log q_n}}$$

and $\sigma(\mu^{\star m})$ will be much smaller than $1/\sqrt{\rho_n}$ as soon as q_n is large enough.

Recall that it takes time of the order $1/\sqrt{\rho}$ for a horocycle to get from $X - X^\rho$ to X^2. Thus for the m found above, there are many, say $c(m)$, standard deviations of $\mu^{\star m}$ around the mean am where $\mathbf{x}_0 u_+([am-c(m), am+c(m)]) \in \mathbf{X} - \mathbf{X}^2$. It follows that if $f \in C_c(\mathbf{X})$ satisfies $\int_{\mathbf{X}} f(\mathbf{x})\omega_X(d\mathbf{x}) = 1$ but f has its support in $\mathbf{X}^2$, we have

$$\liminf_{m\to\infty} \int_{-\infty}^{\infty} f(\mathbf{x}_0 u_+(t))\mu^{\star m}(dt) = 0.$$

This proves that the random walk is not equidistributed. ■

References

[BM00] B. Bekka and M. Mayer. *Ergodic theory and topological dynamics for group actions on homogeneous spaces.* Cambridge University Press, 2000.

[Bre05] E. Breuillard. Local limit theorems and equidistribution of random walks on the heisenberg group. *Geom. Funct. Anal.*, 15(1):35–82, 2005.

[DS84] S. G. Dani and J. Smillie. Uniform distribution of horocycle orbits for Fuchsian groups. *Duke Math. J.*, 51(1):185–194, 1984.

[Fur73] Harry Furstenberg. Boundary theory and stochastic processes on homogeneous spaces. In *Harmonic analysis on homogeneous spaces (Proc. Sympos. Pure Math., Vol. XXVI, Williams Coll., Williamstown, Mass., 1972)*, pages 193–229. Amer. Math. Soc., Providence, R.I., 1973.

[Hed36] G. A. Hedlund. Fuchsian groups and transitive horocycles. *Duke J. Math.*, 2(3):530–542, 1936.

[Hub06] J.H. Hubbard. *Teichmuller theory and applications to geometry, topology and dynamics (vol. 1: Teichmuller theory)*. Matrix Editions, 2006.

[Kin64] A. Kinchin. *Continued fractions*. University of Chicago Press, 1964.

[Rat91a] M. Ratner. Distribution rigidity for unipotent actions on homogeneous spaces. *Bull. Amer. Math. Soc. (N.S.)*, 24(2):321–325, 1991.

[Rat91b] M. Ratner. On Raghunathan's measure conjecture. *Ann. of Math. (2)*, 134(3):545–607, 1991.

[Rat91c] M. Ratner. Raghunathan's topological conjecture and distributions of unipotent flows. *Duke Math. J.*, 63(1):235–280, 1991.

[Rat92] M. Ratner. Raghunathan's conjectures for $\mathrm{SL}(2, \mathbf{R})$. *Israel J. Math.*, 80(1-2):1–31, 1992.